Nguyen Tran Duc Anh - Fresher Data Engineer

Email: ducann02@gmail.com
Phone: (+84) 866 933 124
Website: ducanhntt.github.io
Linkedin: linkedin.com/in/ducanhnt

SUMMARY

I'm an aspiring Data Engineer, passionate about Big Data and Cloud Computing. I enjoy designing automating data pipelines, data processing, ETL/ELT. Even though I'm early in my career, I'm ready to use my knowledge of data technologies to help make business decisions and promote growth.

TECHNICAL SKILLS

Programming languages: Python, Scala, Java

Framework:

- Spark, Kafka, dbt
- HDFS, Hive, HBase
- Airflow, Superset

Database management system: PostgreSQL, MySQL, SQL Server

Tools: Docker, Power BI

Cloud computing: Azure, Snowflake, Databricks

Foreign Language: English - Fluent

Others:

- **Experience in data preprocessing, cleaning, and wrangling**
- **Experience in using version control systems (Git)**
- **Familiarity with Linux command line tools and operations**

PROJECTS

- ### DVD Rental Data Pipeline and Analysis
  July - August 2023
  - **Objective**: *An ELT (Extract - Load - Transform) data pipeline with the DVD Rental database.*
  - **Technologies:** Hadoop HDFS, Spark, Hive, Superset, Postgres, Docker, Apache Airflow.
  - **Extract:** From database (Postgres), ingest to datalake (HDFS).
  - **Load:** Includes 2 main storage system: datalake (HDFS) and Hive (Data warehouse).
  - **Transform:** Preprocess using Scala (with standalone local Spark clusters), and orchestrate using Airflow.
  - **Serving**: Analyze rental trends with BI tools (Superset) and visualize in its own web UI.
  - **Source:** github.com/DucAnhNTT/bigdata-ETL-pipeline
  - **Demo:** youtube.com/playlist?list=PLId1IInL1turLaBYUjjWeEM1z3ZCTMY7d

- ### Sales Data Pipeline with Azure
  Sep - October 2023
  - **Objective**: *A data pipeline using on cloud platform technologies from migrating database on-premise to visualize.*
  - **Technologies:** SQL Server, Data Lake Storage Gen2, Azure Databricks, Data Factory, Power BI, and Synapse Analytics.
  - **Ingestion:** Ingested data from SQL Server into Azure Data Lake Storage Gen2 using Azure Integration Runtime.
  - **ETL Processes:** Databricks is employed for the ETL (Extract, Transform, Load) processes use for transforming the data
  - **Data Modeling and Orchestration:** Azure Data Factory is used for orchestrating the data pipeline workflows.
  - **Data Visualization:** Power BI is utilized for data visualization.
  - **Data Management and Serving:** Lastly, Synapse Analytics acts as Data Warehouse and serving data for BI.
  - **Source:** github.com/DucAnhNTT/azure-data-pipeline-sales
  - **Demo**: youtube.com/playlist?list=PLId1IInL1tur3w-5b9-SY1AvyH8lZw7IA

- **ETL Project: Airflow, Soda, Snowflake Integration** *Sep - October 2023*
    - **Objective**: *An ELT data pipeline utilizing Apache Airflow, Soda for data quality checks, and Snowflake for secure data storage.*
    - **Technologies:** Apache Airflow, Soda (Data quality framework), Snowflake (Cloud-based data warehouse), Astro CLI, Docker.
    - **Project Content:**
        - Developed a data pipeline with Airflow DAGs for ETL workflows.
        - Integrated Soda for automated data quality checks.
        - Utilized Snowflake for flexible and scalable data storage.
    - **Source:** github.com/DucAnhNTT/bigdata-ETL-pipeline
- **Movie Recommendation with Azure** *July - August 2023*
    - **Objective**: *Hands-on experience building an end-to-end data pipeline using Azure services and implementing a movie recommendation system with collaborative filtering and PySpark ML.*
    - **Technologies:** Azure Blob Storage, Azure Databricks, Azure DataFactory, Azure Logic App, Azure AD, Key Vault.
    - **How it Works:**
        - **Data Ingestion:** Movielens dataset (25M records) ingested into Azure Blob Storage. Azure Storage Account created for secure storage.
        - **Data Flow:** Raw data transformed using Azure Databricks, ETL orchestrated with Azure DataFactory. Key Vault used for secure identity management.
        - **Recommendation:** After apply algorithm take out result from Azure Databricks, then via Azure Logic App send to user
    - **Source:** github.com/DucAnhNTT/movie-recom-pipeline-azure
    - **Demo:** youtube.com/playlist?list=PLId1IInL1tup-76xOBJcgUi2A5fy2eGTc

EDUCATION

- Opening University Ho Chi Minh City *Oct. 2021 – Present*
  Bachelor of Computer Science
    - **Revelant Coursework:** Databases, Database Management Systems, Data Mining, Data Structures and Algorithms, Probability and Statistics, Discrete Math.